ФЕДЕРАЛЬНОЕ АГЕНТСТВО ВОЗДУШНОГО ТРАНСПОРТА
(РОСАВИАЦИЯ)

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ГРАЖДАНСКОЙ АВИАЦИИ» (МГТУ ГА)

ФАКУЛЬТЕТ Авиационных систем и комплексов

КАФЕДРА Электротехники и авиационного электрооборудования

Направление подготовки 09.06.01 Информатика и вычислительная техника
(код и наименование направления подготовки)

Направленность 05.13.06 Автоматизация и управление технологическими
(наименование направленности)
процессами и производствами (транспорт)

НАУЧНО-КВАЛИФИКАЦИОННАЯ РАБОТА

Тема **Методы и средства автоматизации процесса патентного поиска при разработке перспективной авиационной техники**

Обучающийся: Григорьев Д.В. (Ф.И.О.) _____ (Подпись)

Научный руководитель: д.т.н., профессор Халютин С.П. (уч.степень, уч.звание, Ф.И.О.) _____ (Подпись)

Рецензенты: д.т.н., профессор Старостин И.Е. (уч.степень, уч.звание, Ф.И.О.) _____ (Подпись)

к.т.н., доцент Гатовский В.А. (уч.степень, уч.звание, Ф.И.О.) _____ (Подпись)

Работа допущена к защите:

Заведующий кафедрой д.т.н., профессор Халютин С.П. (уч.степень, уч.звание, Ф.И.О.) _____ (Подпись)

МОСКВА 2023

Актуальность темы.

В современных условиях при разработке перспективной авиационной техники (АТ) предъявляются повышенные требования к скорости получения новых решений, определения альтернативных вариантов исполнения. Создание конкурентоспособной АТ возможно только с применением лучших решений, поэтому необходимо использовать наработки сторонних организаций, представленных в виде патентов, научных статей и других публикаций.

Осуществление поиска необходимых документов по широким технологическим ландшафтам, как основной способ получения подобной информации, крайне затруднен. При проведении патентных исследований (ПИ) по технологическим ландшафтам характерно наличие большого количества неструктурированной информации о документах, удовлетворяющих критериям отбора при поиске по ключевым словам. Технологические ландшафты, как правило, описываются объемными облаками ключевых слов, что и обусловливает большое количество результатов отбора.

Исходя из этого, тема работы, направленная на исследование методов и средств автоматизации патентного поиска при разработке авиационной техники, является весьма актуальной в процессе эксплуатации бортового оборудования.

Объект исследования: патентные исследования, проводимые при разработке авиационной техники.

Предмет исследования: методы проведения патентного поиска с использованием средств автоматизации.

Цель исследования: повышение эффективности проведения патентных исследований при разработке перспективной авиационной техники за счет использования методов автоматизации поиска.

Задачи исследования:

- анализ предметной области процессов проведения патентных исследований;
- разработка функциональной модели процесса проведения ПИ.
- разработка концептуальной модели системы АСПП.

- проработка вопроса использования технологии построения структурно-функциональных технологических схем для описания объектов поиска.
- анализ предметной области «СЭС ЛА», выделение характеристических признаков объектов СЭС ЛА для поиска.
- разработка методики формирования цифровых поисковых образов объектов в разрезе автоматизированного формулирования поисковой задачи в АСПП.
- разработка методики формирования семантического ядра АСПП для проведения автоматизированного поиска объектов (на примере СЭС ЛА)
- анализ методов тематического моделирования для кластеризации результатов ПИ;
- разработка методики группировки (кластеризации) результатов поиска.
- разработка компьютерной модели прототипа АСПП;
- проведение экспериментальных исследований разработанной методики;
- формирование рекомендаций по применению предложенных решений по автоматизации процесса патентного поиска.

Методология и методы исследования. Для получения основных результатов диссертационной работы использованы базовые методы математического анализа, методы системного анализа, методы теории вероятностей. Также, в работе использовалось компьютерное моделирования в пакете прикладных программ.

1. Патентные исследования в области разработки авиационной техники

Одним из значимых и эффективных методов прогнозирования развития конструкций, технологических процессов и методов производства авиационной техники (АТ) на среднесрочную перспективу являются патентные исследования. Под патентными исследованиями понимаются исследования технического уровня и тенденции развития объектов техники, их патентоспособности и патентной чистоты на основе патентной и другой научно-технической информации. Описание изобретения, в котором отражаются конкретные инженерные решения, обладает значительными преимуществами по сравнению с другими информационными источниками. Массив заявок и описаний изобретений

характеризует тенденцию научно-технического прогресса, по нему можно получить не только качественную, но и количественную оценку наиболее перспективных направлений развития АТ.

Временной интервал развития изобретений от экспериментов в НИИ и лабораториях, в которых они зародились, до рынка сбыта, где они выступают в материальной форме в виде готовых изделий, обычно составляет в авиационной отрасли 10-15 лет, поэтому наибольшее значение патентная информация приобретает для среднесрочного прогнозирования, хотя может быть использована и для краткосрочных прогнозов. Техническое направление, разработанное сегодня исследователями, и ставшее для них пройденным этапом, остаётся актуальным для производственников, на протяжении10-15 лет.

При выполнении научно-исследовательской работы (НИР) патентные исследования предусмотрены в техническом задании (ТЗ), в том числе в отношении результатов ПИ, а также ответственности сторон за последствия, вызванные выполнением их в ограниченном объеме или отказом от использования их результатов.

Необходимость проведения ПИ при выполнении составных частей работ или при разработке комплектующих изделий, материалов, технологии, осуществляемых по исходному техническому заданию, определяет Головной исполнитель работы в соответствии с ТЗ.

Результаты ПИ отражаются не только в отчете о патентных исследованиях (ОПИ), но и в технических условиях и стандартах на разработанную продукцию, в документации, связанной с оценкой технического уровня и качества продукции.

ПИ в полном объеме в соответствии с ГОСТ Р 15.011 - 96 должны проводиться на начальной стадии НИР, а в дальнейшем на всех стадиях научно-исследовательских, опытно-конструкторских работ научно-исследовательских, опытно-конструкторских работ (НИОКР), связанных с созданием, производством, реализацией, совершенствованием и использованием продукции производственного назначения, рекомендуется дополнять исследования изучением всех новых материалов. Это определяется тем, что объектами

патентной охраны могут быть как сами изделия во всех аспектах их исполнения (схемы, конструкции, технологии изготовления и т.п.), так и методы (способы) использования при эксплуатации (способы измерений, регистрации, обработки информации и т.п.).

Проведение ПИ включает:

- определение задач и разработку задания на проведение ПИ;
- определение требований к поиску патентной и другой документации;
- поиск и отбор патентной и другой документации и оформление отчета о поиске;
- систематизацию и анализ отобранной документации, подготовку выводов и рекомендаций;
- оформление результатов исследований в виде отчета.

Информация по выбору стран и глубины поиска по видам ПИ представлена в таблице 1.

Вид патентных исследований	Страны	Ретроспектива
Уровень техники	Обязательно страны в которых традиционно ведутся разработки в области космической техники: РФ, США, страны ЕС, КНР, Япония, Индия	достаточно 10 лет*
Патентная чистота	РФ и страны, в которых планируется реализация продукции	25 лет** (с учетом возможности продления срока действия патента)
Патентоспособность	без ограничения	без ограничения

** Отраслевые рекомендации – 25 лет*
***В Российской Федерации в отношении изобретений глубина поиска составляет 20 лет*

Таблица 1. Выбор стран и глубины поиска по видам ПИ.

При поиске патентной информации используют электронные ресурсы:

- Яндекс. Патенты, созданные при содействии Федеральной службы по интеллектуальной собственности (Роспатента).

Отраслевые методические рекомендации не рекомендуют использовать данный ресурс, но для полноты картины патентного поиска, как дополнительное средство поиска, БД может быть использована для глобального поиска, так как обладает некоторыми преимуществами, такими как простота и доступность.

- Федеральное государственное бюджетное учреждение «Федеральный институт промышленной собственности» ФИПС База данных российских и иностранных изобретений https://new.fips.ru/iiss/.
- База данных авторских свидетельств СССР http://patentdb.su/.
- Цифровая платформа поиска патентной информации и средств индивидуализации и сервис поиска патентной информации (ИС «Поисковая платформа», запущена в 2022 году) https://searchplatform.rospatent.gov.ru/.

На разработанной поисковой платформе реализован поиск сведений по мировому фонду изобретений и полезных моделей, включая российские массивы и массивы стран СНГ, многоязычный полнотекстовый и атрибутивный поиск на основных европейских языках, поиск на основе патентных классификаторов, поиск с использованием искусственного интеллекта, поиск по химическим формулам, генетическим последовательностям и др. Для разработчиков реализован программный интерфейс API.

Также на платформе размещены аналитические сервисы, которые позволят проводить мониторинг показателей сферы интеллектуальной собственности.

Завершаются работы для обеспечения поиска на цифровой платформе сведений по товарным знакам и промышленным образцам. По мере завершения работ доступ к этим объектам появится на платформе.

Источниками научно-технической информации являются:

- реферативная информация о последних достижениях науки и техники;
- реферативная информация о патентах зарубежных стран;
- полные описания изобретений к авторским свидетельствам и патентам;
- отчеты о НИР, ОКР, ПКР;
- диссертации;

• материалы профильных конференций;

• материалы непубличных источников;

• доклады о перспективах развития исследуемой области техники;

• отчеты о патентных исследованиях;

• официальные нормативные документы;

• стандарты;

• технические условия;

• конъюнктурно-экономическая информация (рекламные проспекты, каталоги, справочники и т.п.);

• другая доступная, достоверная научно-техническая литература и информация.

Порядок проведения ПИ состоит из следующих этапов:

• поиск и отбор патентной и другой документации в соответствии с утвержденным регламентом проведения патентного поиска и оформление отчета о поиске;

• систематизацию и анализ отобранной документации;

• обоснование решений задач патентных исследований;

• обоснование предложений по дальнейшей деятельности организации, подготовка выводов и рекомендаций при определении уровня техники;

• оформление результатов исследований в виде отчета о патентных исследованиях.

Проведение и оформление отчета регламентировано нормативными документами:

• «ГОСТ Р 15.011 - 96. Государственный стандарт Российской Федерации. Система разработки и постановки продукции на производство. Патентные исследования. Содержание и порядок проведения»;

• «ГОСТ 7.32-2017. Межгосударственный стандарт. Система стандартов по информации, библиотечному и издательскому делу. Отчет о научно-исследовательской работе. Структура и правила оформления».

При выполнении аналитической части отчета все приведенные в отчете выводы должны быть обоснованы проведенным анализом.

Заключение отчета о поиске должно содержать:

1. Обобщенные выводы по результатам проведенных ПИ, например, это могут быть:
   - основные тенденции развития в данной области;
   - ведущие страны в данной области;
   - наибольшее количество изобретений было заявлено и в каких годах;
   - какие фирмы являются ведущими – патентовладельцами
2. Оценку состояния выполнения работы, составной частью которой являются ПИ (например, НИР и ОКР), в свете соответствия его требованиям к конечным результатам работы, целям, планам, программам, перспективам деятельности предприятия (организации);
3. Предложения по использованию результатов ПИ для совершенствования научно-технической, производственной продукции, услуг и развития деятельности предприятия (организации).

2. Моделирование автоматизированной системы патентного поиска

Для устранения существующих недостатков технологии проведения патентного поиска при создании новой АТ необходимо проведение более обширных углубленных патентных исследований в требуемой области с поддержкой специализированных инструментальных средств (автоматизированных систем патентного поиска (АСПП). Структура комплекса представлена в виде диаграммы компонентов в нотации UML на Рисунке 2.

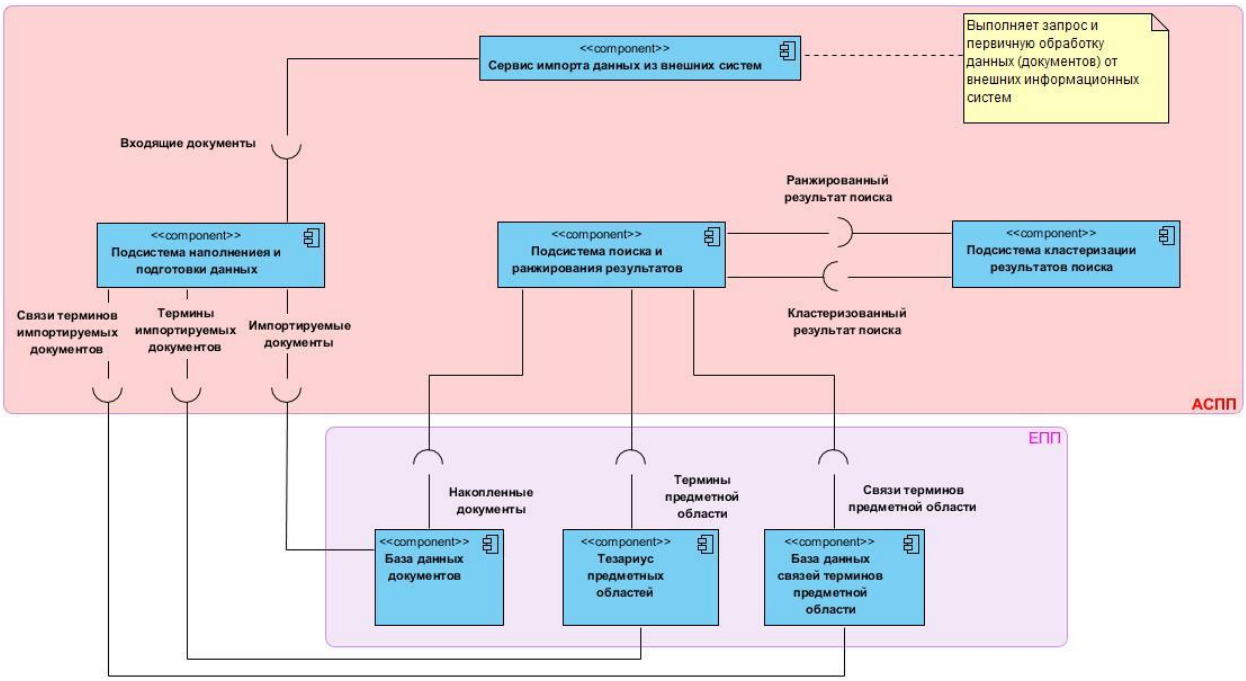


Рисунок 2. Диаграмма компонентов АСПП.

На данной схеме представлен состав и взаимодействие подсистем комплекса на уровне данных.

Сервис импорта данных из внешних систем – подсистема, осуществляющая в автоматическом режиме сбор данных (документов) из подключенных источников и передачу их подсистеме наполнения и подготовки данных.

Подсистема наполнения и подготовки данных - осуществляет занесение документов в базу данных единого поискового пространства (ЕПП), производится индексация содержимого документов с заполнением таблиц ключевых слов, расчет частоты их присутствия в документе и установление весовых коэффициентов значимости (ВКЗ) ключевых слов в рамках данного документа.

База данных документов в ЕПП – информационный ресурс, содержащий данные документов.

База данных связей терминов предметной области – база данных, содержащая состав и весовые коэффициенты значимости ключевых слов в документах по предметным областям.

Тезаурус предметной области – таблица, содержащая информацию о принадлежности ключевых слов предметным областям.

Подсистема поиска и ранжирования результатов – осуществляет выборку документов из хранилища ЕПП в соответствии с запросом пользователя и ранжирование результатов поиска на основании ВКЗ ключевых слов в облаке поисковых терминов.

Подсистема кластеризации результатов поиска – функциональная подсистема, предназначенная для выполнения процесса кластеризации (группировки) найденных документов, в ходе которого определяется семантическая близость документов на основании находящихся в них ключевых слов и их ВКЗ.

Для реализации взаимодействия пользователя с данным программным комплексом был разработан следующий сценарий работы, EPC диаграмма которого приведена на Рисунке 3.

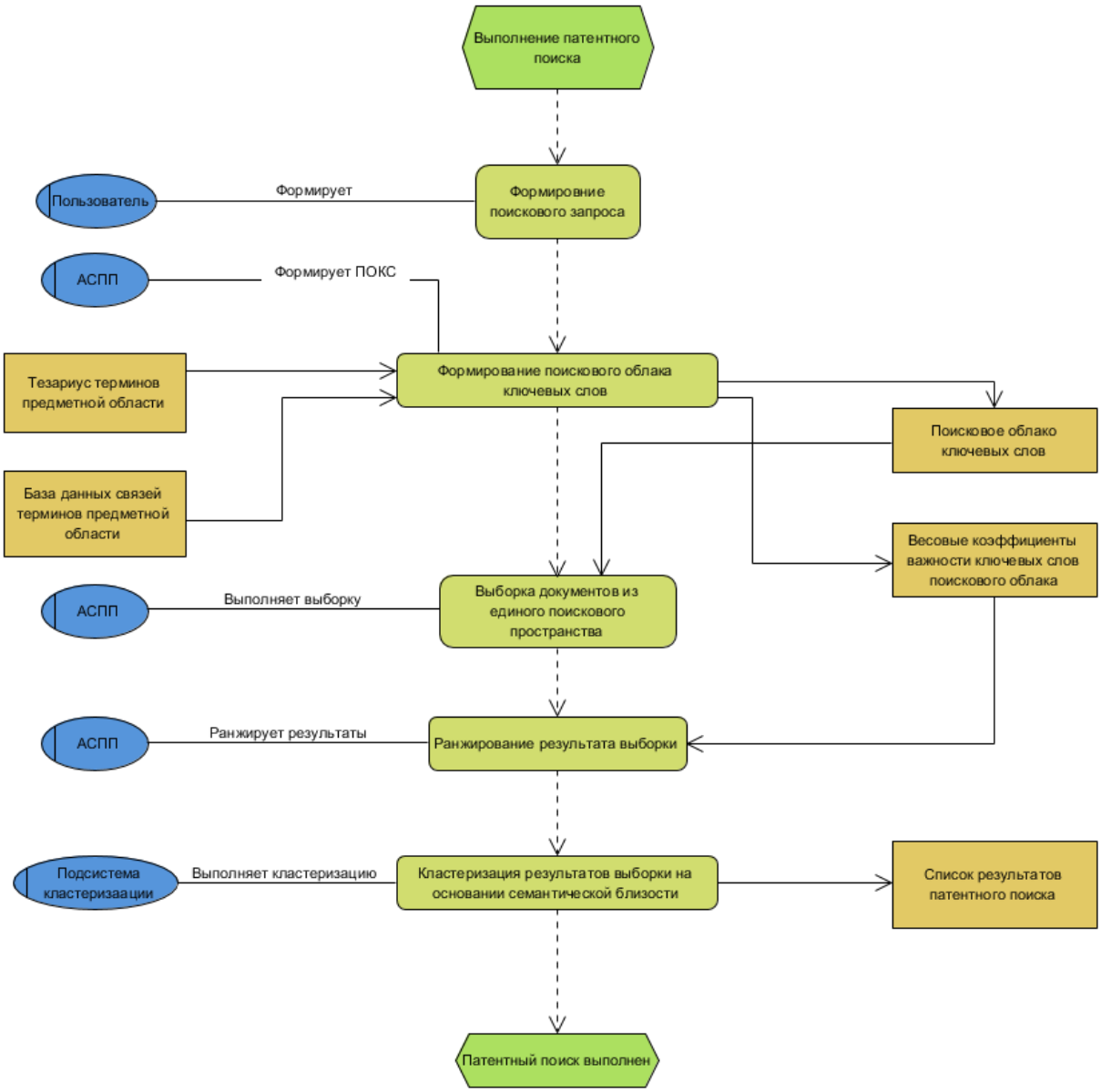


Рисунок 3. EPC диаграмма сценария работы пользователя в АСПП

Функциональные возможности системы представлены в виде диаграммы вариантов использования, которая приведена на Рисунке 4.

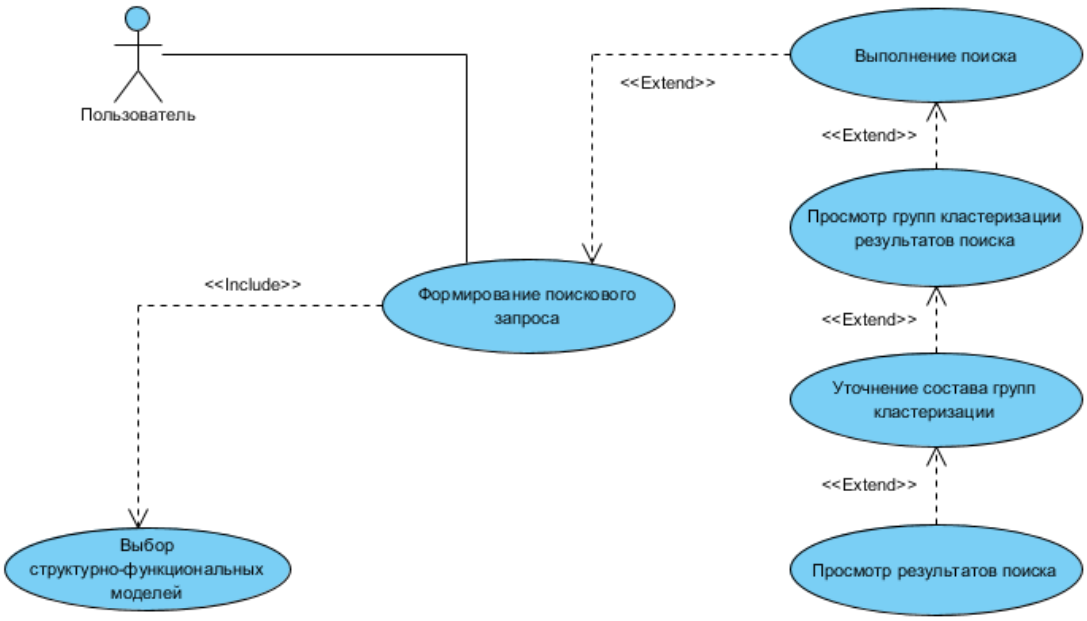


Рисунок 4. Диаграмма варианта использования АСПП.

3. Моделирование предметной области СЭС ЛА

Проведен анализ классификации и на его основе разработана модель предметной области СЭС ЛА, которая представлена в виде диаграммы компонентов в нотации UML на рисунке 7

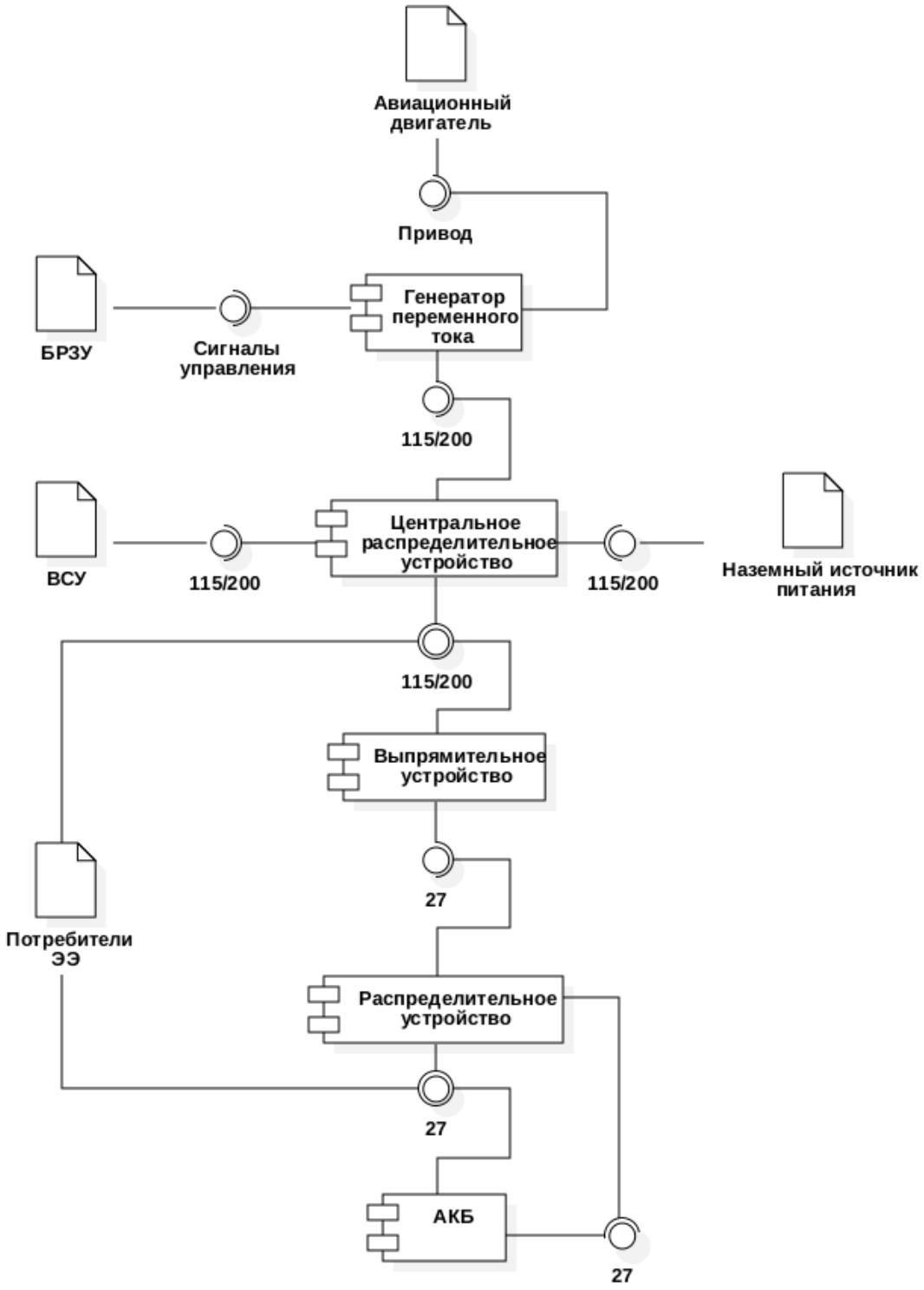


Рисунок 7. Диаграмма компонентов СЭС ЛА.

На данной диаграмме представлены основные компоненты системы электроснабжения ЛА и взаимосвязи между ними с использованием формализации нотации UML. Здесь в виде артефактов обозначены внешние сущности для рассматриваемой системы: авиационный двигатель, вспомогательная силовая установка (ВСУ), наземный источник питания, блок

регулирования защиты и управления (БРЗУ), потребители электроэнергии. В виде элементов схемы представлены объекты генерирующей, распределительной и преобразовательной подсистем. На связях указаны основные параметры взаимодействия между объектами.

Данная модель позволяет разработать схему базы данных предметной области, которая станет ядром хранения информации о ЦПО объектов поиска. Классификация объектов, проведенная ранее, позволяет определить ключевые характеристические атрибуты классов элементов СЭС ЛА.

4. Подсистема группировки результатов ПИ

Предлагаемая подсистема семантического анализа коллекции текстовых документов предназначена для разбиения набора документов на отдельные группы, каждая из которой соответствует определенной теме. Иными словами, подсистема должна решать задачу тематического моделирования на входящем множестве текстовых документов.

Построение тематической модели может рассматриваться как задача одновременной кластеризации документов и слов по одному и тому же множеству кластеров, которые и отражают темы поисковых запросов. Тема – это результат би-кластеризации, то есть одновременной кластеризации и слов, и документов по их семантической близости. Под определением темы в контексте данной задачи обычно понимают набор близких по смыслу слов, которые наиболее часто появляются вместе в документах общей направленности.

По структуре система кластеризации состоит из нескольких модулей (рисунок 8):

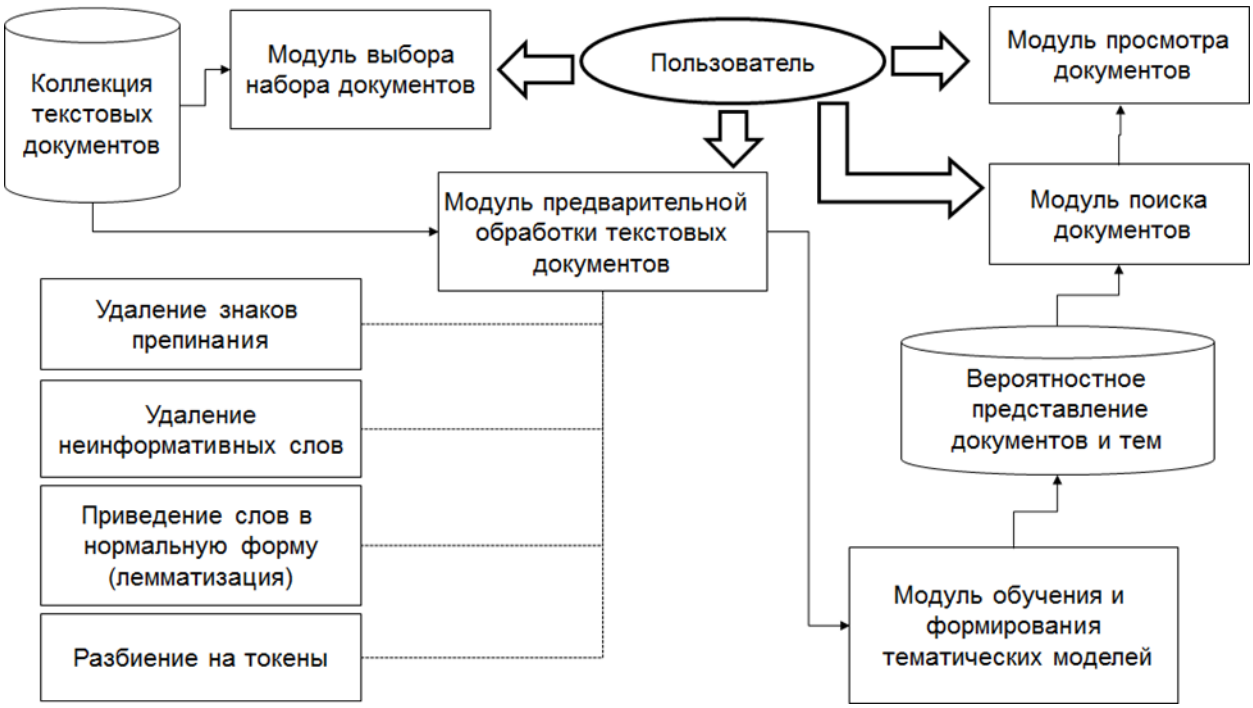


Рисунок 8. Структурная схема подсистемы кластеризации

Модуль предварительной обработки текстовых документов отвечает за приведение текстовых данных к удобному для последующей обработки формату, а также за проведение других предварительных операций с текстами, позволяющими значительно улучшить итоговое качество работы системы. Основные функции, реализованные в данном модуле – это:

- удаление знаков препинания - с точки зрения выбранной для решения задачи математической модели знаки препинания не несут в себе никакой смысловой информации и, следовательно, не должны принимать участие в ее построении;
- удаление неинформативных слов (стоп-слов) - в текстах на любых языках встречается большое количество слов, мало интересующих исследователя с точки зрения выделения тематики документов в коллекции; такие слова называются стоп-словами (например, слова если, как, что, он, когда и другие); их присутствие в текстах документов не только замедляет их обработку, но и негативно сказывается на интерпретируемости (понятности пользователю) результирующих тем, и, поэтому, их необходимо отбрасывать на этапе предварительной обработки;

- приведение слов к нормальной форме [8] (лемматизация) - крайне важная процедура, смысл которой заключается в том, чтобы различные формы одного и того же слова (например, слово генератор в других падежах – генератора, генератору, генератором и т.д.) рассматривались программой как одно слово. Реализация данного функционала положительно сказывается как на скорости работы программы, так и на качестве итоговой модели.
- разбиение на термы – процедура, задача которой заключается в представлении текста в виде отдельных составляющих, термов. В качестве термов могут выступать отдельные слова, словосочетания, предложения и др.

Модуль обучения и формирования тематических моделей реализует основную функциональную часть системы, а именно выделение набора тем из коллекции документов и построение матриц вероятностей $\varphi_{\omega t}$ и θ_{td},, соответствующих выбранной тематической модели.

5. Метод тематического моделирования

Пусть $D = \{d_1, \dots, d_n\}$– множество (коллекция) текстовых документов, $W = \{w_1, \dots, w_m\}$ – множество (словарь) всех употребляемых в них термов. Термами [9] могут быть слова, нормальные формы слов, словосочетания, предложения и др., в зависимости от того, какие методы предварительной обработки были выполнены.

Каждый документ $d \in D$ представляет собой последовательность термов n_d термов $\omega_1 \dots \omega_{n_d}$из словаря W.

Гипотеза о существовании тем [10] гласит, что каждое вхождение терма ω в документ d связано с некоторой темой t из заданного конечного множества T. Также предполагаем, что число тем много меньше, чем число документов и термов.

Коллекция документов представляет собой последовательность троек

$$D = \{(W_i, d_i, T_i) | \; i = 1, \dots, n\},$$ где

d_i – документ из множества D,

W_i – вектор термов документа d_i,

T_i –вектор тем документа d_i,

Термы ω_i и документы d_i являются наблюдаемыми переменными, темы t_iнеизвестны и являются латентными (скрытыми) переменными.

Порядок термов в документах не важен для выявления тематики. Это предположение называется гипотезой «мешка слов» [11]. Порядок документов в коллекции также не имеет значения.

Согласно гипотезе об условной независимости [10], появление слов в документе d по теме t зависит от темы, но не зависит от документа d и описывается общим для всех документов распределением $p(\omega \mid t)$:

$$p(\omega \mid d, t) = p(\omega \mid t)$$

Распределение термов в документе $p(\omega \mid d)$ описывается *вероятностной смесью* распределений термов в темах $\varphi_{\omega t} = \; p(\omega \mid t)$ с весами $\theta_{td} = \; p(t \mid d)$:

$$p(\omega \mid d) = \sum_{t \in T} p(\omega \mid t, d)\, p(t \mid d) = \sum_{t \in T} p(\omega \mid t)\, p(t \mid d) = \sum_{t \in T} \varphi_{\omega t} \theta_{td}\,,$$ где

$p(\omega \mid d)$ — вероятность появления терма ω в документе d, $p(\omega \mid t)$ — вероятность появления терма ω в документе c темой t, $p(t \mid d)$ — вероятность присутствия темы t в документе d.

Задача тематического моделирования [12] – обратная задаче порождения коллекции документов, т. е. по заданной коллекции документов D требуется найти множество тем T, распределения $p(\omega \mid t)$ для всех тем $t \in T$ и распределения $p(t \mid d)$ для всех документов $d \in D$ и параметры распределения термов в темах и тем в документах соответственно φ_{wt} и θ_{td}.

Тематическая модель латентного размещения Дирихле (LDA) [10] основана на разложении документов при дополнительном предположении, что векторы

документов $\theta_d = (\theta_{td}) \in R^{|T|}$ и векторы тем $\varphi_t = (\varphi_{wt}) \in R^{|W|}$ порождаются распределениями Дирихле с параметрами $\alpha \in R^{|T|}$ и $\beta \in R^{|W|}$ соответственно:

$$Dir(\theta_d;\, \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1},\ \alpha_t > 0,\ \alpha_0 = \sum_t \alpha_t,\ \theta_{td} > 0,\ \sum_t \theta_{td} = 1;$$

$$Dir(\varphi_t;\, \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \varphi_{wt}^{\beta_w - 1},\ \beta_w > 0,\ \beta_0 = \sum_w \beta_w,\ \varphi_{wt} > 0, \sum_w \varphi_{wt} = 1.$$

где Γ (z) – гамма-функция;

$R^{|W|}$ –пространство термов;

$R^{|T|}$ - пространство тем.

Математическое ожидание и дисперсия t-й координаты вектора θ_d [7] равны, соответственно,

$$E\theta_{td} = \int \theta_{td} Dir(\theta_d; \alpha) d\theta_d = \frac{\alpha_t}{\alpha_0}, \qquad D\theta_{td} = \frac{\alpha_t(\alpha_0 - \alpha_t)}{\alpha_0^2(\alpha_0 + 1)}.$$

Векторный параметр α определяет [13] степень разреженности векторов θ_d, пороржданемых распределением $Dir(\theta;\, \alpha)$. Если $\alpha_t = 1$ для всех t, то распределение Дирихле переходит в равномерное. Чем больше α_0, тем меньше дисперсия, и тем сильнее векторы θ_d концентрируются вокруг вектора математического ожидания $E\theta_d$. Чем меньше α_t, тем сильнее значения θ_{td} концентрируются вокруг нуля. Чем меньше α_0, тем более разрежен вектор θ_d. Поэтому α_t называют параметрами контраста [7].

Распределение Дирихле – достаточно широкое параметрическое семейство распределений на единичном симплексе [7], которое описывает как разреженные, так и сконцентрированные дискретные распределения.

Модель LDA подходит для описания кластерных структур [14]. Чем меньше значения гиперпараметров α и β, тем сильнее разрежено распределение Дирихле, и тем дальше отстоят друг от друга порождаемые им векторы. Чем меньше α_0, тем сильнее различаются документы θ_d. Чем меньше β_0, тем сильнее различаются темы φ_t. Векторы $\varphi_t = p(w|t)$ в пространстве термов $R^{|W|}$ представляют центры тематических кластеров. Элементами кластеров являются векторы документов с

эмпирическими распределениями $\hat{p}(w \mid d, t)$. Чем меньше гиперпараметры β, тем больше межкластерные расстояния по сравнению с внутрикластерными. Таким образом, гиперпараметры позволяют моделировать тематические кластерные структуры различной степени выраженности [15].

Распределение Дирихле также является сопряженным к мультиномиальному, что упрощает вывод апостериорных оценок вероятностей θ_{td} и φ_{wt} [7].

6. Модифицированный метод латентного размещения Дирихле для формирования ЦПО объектов

Учитывая возможности разбиения текстов документов на составляющие смысловые термы и тематической группировки по методу Дирихле предлагается на этапе формирования би-грамм осуществлять отбор и семантическое ранжирование полученных векторов термов.

Это достигается осуществлением обучения модели на заранее подготовленных текстах, специально отобранных экспертами и относящихся к теме исследуемой предметной области. На этапе формирования би-грамм формируется облако ключевых понятий имеющих отношение к теме предметной области. Механизм вероятностного распределения термов (в нашем случае биграмм) по темам позволяет установить весовые коэффициенты значимости конкретной биграммы и достигнуть семантического ранжирования. Данное облако ключевых термов с весовыми коэффициентами значимости составляют основу автоматизировано формируемого цифрового поискового образа объектов, семантически связанных с темой исследования.

7. Программная реализация системы

На первом этапе обработки осуществляется импорт данных из файлов, полученных в результате поиска.

In [12]:
```
data = pd.DataFrame(data_dir)
```

In [13]:
```
data
```

Out[13]:

	text	tema
0	Процесс включения синхронного генератора на па...	Автоматизация включения генераторов на паралле...
1	В тех случаях, когда не требуется гальваническ...	Автотрансформаторные ВУ
2	Выпрямители на диодах и тиристорах обладают из...	Активные выпрямители
3	\nВ системах электроснабжения ЛА имеют место к...	Анализ динамики процессов регулирования напряж...
4	Анализ процессов регулирования напряжения гене...	Анализ динамики процессов регулирования напряж...
...	...	...
112	Принцип действия серебряно-цинковых аккумулято...	Устройство принцип действия и основные характе...
113	Химический источник тока (ХИТ) -- устройство, ...	Химические источники тока
114	В цифровых регуляторах напряжения (ЦРН) исполн...	Цифровые регуляторы напряжения
115	Среди электромеханических приводов наиболее эф...	Электромеханический привод
116	Принципиальное отличие процессов электрохимиче...	Электрохимичекие процессы в топливных элементах

117 rows × 2 columns

Следующим этапом идёт очистка текстов, удаление неинформативных слов, знаков препинания и лемматизация текста.

In [20]:
```
train_lst[0]
```

Out[20]:
```
'вертолёт отбор мощность привод генератор правило осуществляться вал главное редуктор несущий винт частота вращение который стабилизироваться высокий точность связь это синх
ронный генератор приводиться вращение непосредственно редуктор двигатель применение привод постоянный скорость учитывать данный обстоятельство вертолёт применять преимуществ
енно первичный система электроснабжение переменный ток первичный система электроснабжение постоянный ток вертолёт применять запуск двигатель осуществляться электростартер ст
руктурный схема типовой система электроснабжение вертолёт привести рис структурный схема система электроснабжение вертолёт первичный система электроснабжение содержать незав
исимый канал генерирование электроэнергия переменный ток каждый который включать свой состав бесконтактный синхронный генератор регулятор напряжение блок защита управление б
зу блок трансформатор ток бтт связь генератор вращаться синхронно один вал редуктор система электроснабжение предусмотреть параллельный работа объединение канал параллельный
работа осуществляться вручную автоматически сигнал бзу помощь контактор параллельный работа кпр равномерный распределение мощность генератор регулятор напряжение подаваться
сигнал пропорциональный разность полный ток нагрузка сигнал поступать блок трансформатор ток оба канал генерирование который соединяться себя дифференциальный схема блок защ
ита управление обеспечивать автоматический включение генератор сеть также защита соответствующий канал генерирование снижение частота аварийный повышение снижение напряжение
обрыв фаза вид короткий замыкание защита аварийный повышение частота переменный ток вертолёт предусматриваться связь непосредственный привод генератор редуктор несущий винт
вид авария невозможный питание приёмник трехфазный ток напряжение однофазный приёмник использоваться соответственно понижать трансформатор который являться основный резервны
й переключение питание приёмник короткий замыкание обрыв фаза снижение напряжение основный трансформатор резервный осуществляться помощь автомат переключение преобразователь
апп получение постоянный ток напряжение система электроснабжение использоваться выпрямительный устройство нормальный работа система электроснабжение объединяться параллельны
й работа характерный особенность рассматривать система электроснабжение являться высокий степень резервирование питание распределительный устройство обеспечивать хороший экс
плуатационнотехнический характеристика система например отказ основный однофазный трансформатор приёмник переменный ток получать питание резервный трансформатор отказ аварий
ный источник электромашинный преобразователь тип аналогично осуществляться питание приёмник трехфазный переменный ток напряжение отказ один приёмник автоматически переключат
ься питание исправный помощь контактор случай отказ оба аккумуляторный батарея обеспечивать питание приёмник подключить дмр срабатывать обратный ток изолировать остальной пр
иёмник постоянный ток параллельно аккумуляторный батарея работать также стартергенератор стг который включаться запуск вспомогательный силовой установка'
```

In [21]:
```
# Подсчитаем кол-во слов в предложениях (составление терм-документной матрицы)
vector_ben = CountVectorizer(
    analyzer='word',
    # min_df - частота встречаемого термина < 10
    min_df=10,
    # биграммы, триграммы
    ngram_range=(2, 3),
    stop_words=stopwords.words("russian"))
train_vec_ben = vector_ben.fit_transform(train_lst)
```

In [22]:
```
len(train_lst)
```

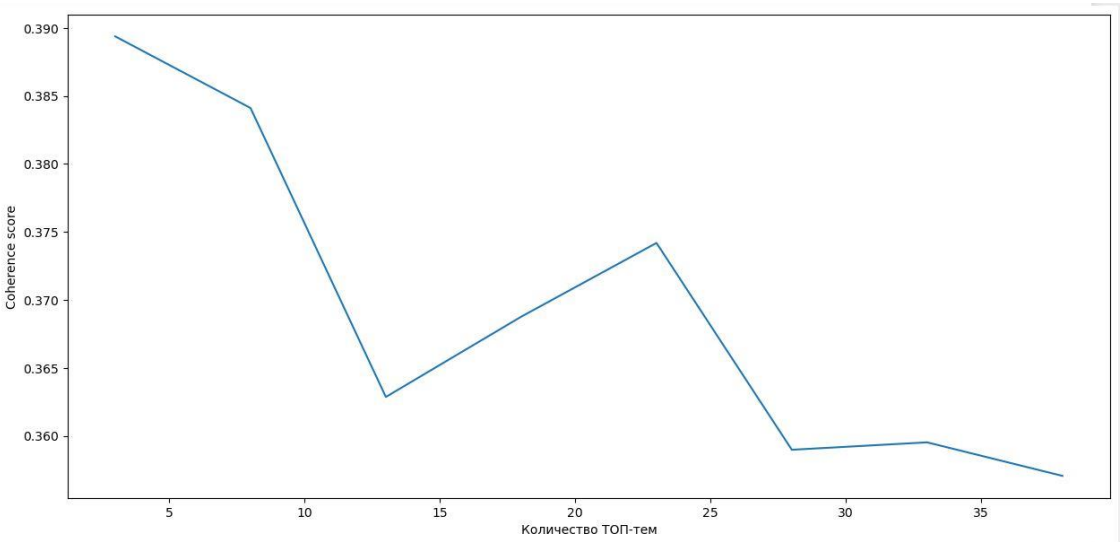
Out[22]:
```
104
```

Затем производится разбиение корпусов текстов в документах на биграммы с получением векторов коэффициентов значимости.

```
In [26]: vector_ben.get_feature_names_out()

Out[26]: array(['аварийный источник', 'аварийный повышение', 'аварийный режим',
       'авиационный двигатель', 'аккумуляторный батарея',
       'активный мощность', 'активный нагрузка', 'активный сопротивление',
       'амплитудный значение', 'аппарат защита', 'блок управление',
       'бортовый сеть', 'векторный диаграмма', 'величина напряжение',
       'величина ток', 'включение генератор', 'возбуждение генератор',
       'вращение генератор', 'вспомогательный силовой',
       'вспомогательный силовой установка', 'вторичный обмотка',
       'вторичный обмотка трансформатор', 'входной напряжение',
       'выпрямительный устройство', 'выпрямить напряжение',
       'выходной напряжение', 'генератор иметь', 'генератор напряжение',
       'генератор параллельный', 'генератор параллельный работа',
       'генератор переменный', 'генератор переменный ток',
       'генератор постоянный', 'генератор постоянный ток',
       'генератор работать', 'генератор равный', 'генератор регулятор',
       'генератор система', 'генератор ток',
       'генерирование электроэнергия', 'действовать значение',
       'допустимый значение', 'задать значение', 'зажим генератор',
       'защита управление', 'значение коэффициент', 'значение напряжение',
       'значение ток', 'изменение напряжение', 'изменение ток',
       'измерительный орган', 'иметь вид', 'иметь место',
       'источник переменный', 'источник переменный ток',
       'источник питание', 'источник постоянный', 'источник ток',
       'источник электроэнергия', 'канал генерирование',
       'канал генерирование электроэнергия', 'качество пример',
       'качество электрический', 'качество электрический энергия',
       'качество электроэнергия', 'короткий замыкание',
       'коэффициент усиление', 'линейный напряжение',
       'максимальный значение', 'момент время', 'мощность генератор',
       'нагрузка генератор', 'нагрузка напряжение', 'напряжение выход',
       'напряжение генератор', 'напряжение зажим', 'напряжение иметь',
       'напряжение источник', 'напряжение который', 'напряжение нагрузка',
       'напряжение напряжение', 'напряжение переменный',
       'напряжение переменный ток', 'напряжение питание',
       'напряжение постоянный', 'напряжение постоянный ток',
       'напряжение привести', 'напряжение равный', 'напряжение рис',
       'напряжение сеть', 'напряжение система', 'напряжение ток',
       'напряжение частота', 'настоящий время', 'номинальный значение',
       'номинальный напряжение', 'нормальный работа', 'нормальный режим',
       'обмотка возбуждение', 'обмотка трансформатор', 'обратный связь',
```

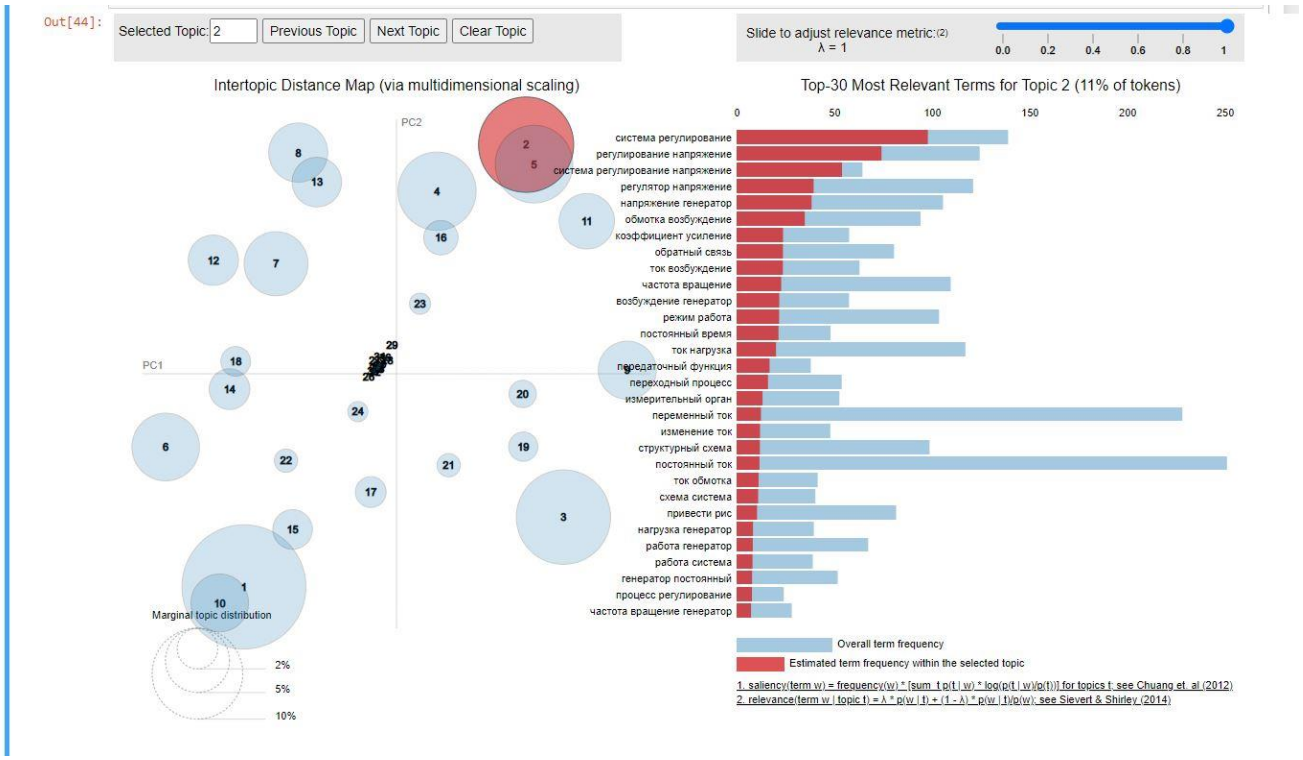
Осуществляем подбор оптимального количества тем для группировки результатов.




```
In [41]: for n_topics, coherence in zip(x, coherence_values):
             print(f"ТОП-тем = {n_topics}, Coherence = {coherence:.3f}")

ТОП-тем = 3, Coherence = 0.389
ТОП-тем = 8, Coherence = 0.384
ТОП-тем = 13, Coherence = 0.363
ТОП-тем = 18, Coherence = 0.369
ТОП-тем = 23, Coherence = 0.374
ТОП-тем = 28, Coherence = 0.359
ТОП-тем = 33, Coherence = 0.360
ТОП-тем = 38, Coherence = 0.357
```

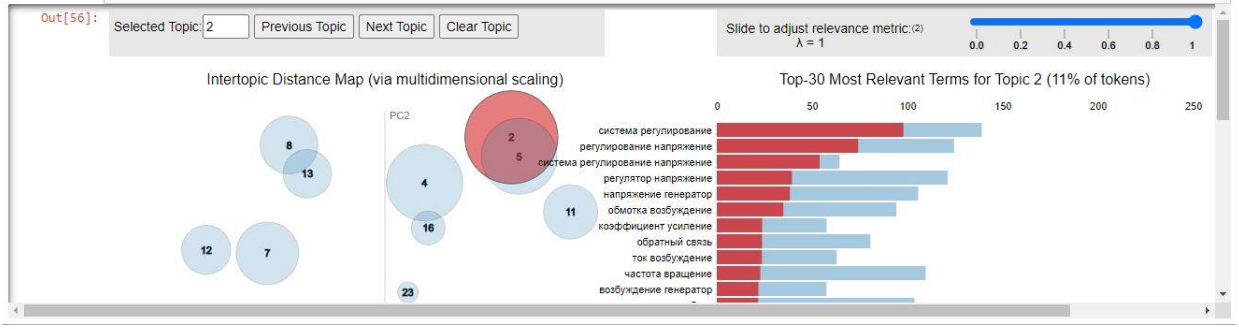
Ниже приведена графическая интерпретация группировки документов по определённым на предыдущем этапе темам.



И представление семантически значимых терминов в документах с пересечением тем.

```
In [30]: len(lda_model_ben.components_)
Out[30]: 10

In [56]: pyLDAvis.enable_notebook()
         panel = pyLDAvis.lda_model.prepare(lda_model_ben,
                                            test_vec_ben,
                                            vector_ben,
                                            mds='tsne')
         panel
```



Группировка тем осуществляется по критерию, формируемому при определении степени когерентности тем.

Тема называется **когерентной** (согласованной), если её топ-N слова встречаются вместе чаще, чем можно было бы ожидать от случайного распределения (топ-N наиболее вероятных слов для конкретной темы). Высококогерентная тема содержит семантически связанные между собой слова. Метрика коррелирует с экспертными оценками.

```
In [36]: def get_coherence_mean(model, texts, n_top_words=20):
             """Получение средней оценки когерентности"""

             # кол-во тем
             topics = model.components_

             # получение токенов
             texts = [[word for word in doc.split()] for doc in texts]
             # создание словаря с использованием gensim
             dictionary = corpora.Dictionary(texts)

             # Корпус на основе матрицы подсчета слов
             corpus = [dictionary.doc2bow(text) for text in texts]

             feature_names = [dictionary[i] for i in range(len(dictionary))]

             # Получение ТОП-слов для каждой темы
             top_words = []
             for topic in topics:
                 top_words.append(
                     [feature_names[i] for i in topic.argsort()[:-n_top_words - 1:-1]])

             coherence_model = CoherenceModel(topics=top_words,
                                              texts=texts,
                                              dictionary=dictionary,
                                              coherence='c_v')
             coherence = coherence_model.get_coherence()
             return coherence

In [37]: get_coherence_mean(lda_model_ben, test_lst)
Out[37]: 0.34060134986165086
```

На построенной и обученной модели возможно выполнение запросов с целью получения документов, семантически связанных с ключевыми словами запроса.

Сформируем запрос системе на выдачу документов, имеющих отношение к «синхронный генератор». Ниже приведен результат обработки запроса с проведением анализа семантической близости документов.

```
In [35]: data['top_topic'] = data.text.transform(predict_topic)

In [36]: data[:5]
```

Out[36]:

	text	tema	top_topic
0	Процесс включения синхронного генератора на па...	Автоматизация включения генераторов на паралле...	2
1	В тех случаях, когда не требуется гальваническ...	Автотрансформаторные ВУ	34
2	Выпрямители на диодах и тиристорах обладают из...	Активные выпрямители	28
3	\nВ системах электроснабжения ЛА имеют место к...	Анализ динамики процессов регулирования напряж...	5
4	Анализ процессов регулирования напряжения гене...	Анализ динамики процессов регулирования напряж...	5

```
In [37]: predict_topic('синхронный генератор')
Out[37]: 30
```

В результате имеем рекомендованный документ (с индексом 30) из поисковой совокупности, наиболее связанный семантически с текстом запроса.

Таким образом, система выдает рекомендацию и осуществляет отбор документа из поисковой совокупности по запрошенному понятию, которое не является буквальным ключевым словом, содержащемся в документе.

Данный подход позволяет осуществлять отбор документов по свободно лингвистически формируемым запросам.

Ограничением текущей реализации является длина запроса, которая не может превышать 3 слов (биграммы и триграммы). Данное ограничение возможно снять при обучении системы по значительно большему количеству эталонных документов, имеющих семантическое отношение к предметной области.

ЗАКЛЮЧЕНИЕ

Основными результатами, полученными в ходе выполнения научно-квалификационной работы, являются:

1. анализ предметной области процессов проведения патентных исследований;
2. разработка функциональной модели процесса проведения ПИ.
3. разработка концептуальной модели системы АСПП.
4. Рассмотрен вопрос использования технологии построения структурно-функциональных технологических схем для описания объектов поиска.

5. анализ предметной области «СЭС ЛА», выделены характеристические признаки объектов СЭС ЛА для поиска.
6. разработка методики формирования ЦПО объектов в разрезе автоматизированного формулирования поисковой задачи в АСПП.
7. разработана методика формирования семантического ядра АСПП для проведения автоматизированного поиска объектов (на примере СЭС ЛА) с использованием модифицированного метода латентного размещения Дирихле.
8. анализ методов тематического моделирования для кластеризации результатов ПИ;
9. разработка методики группировки (кластеризации) результатов поиска.
10. разработана компьютерная модель системы тематического моделирования результатов поиска с использованием языка программирования Python.
11. проведены экспериментальные исследования разработанной методики;
12. формирование рекомендаций по применению предложенных решений по автоматизации процесса патентного поиска.